UCRL-ID-146419

FY02 I/O Integration Blueprint

K. C. Cupps
M. R. Gary
K. J. Fitzgerald
T. M. Quinn
D. P. Wiltzius
K. J. Minuzzo
M. K. Seager
T. T.McLarty

November 16, 2001

U.S. Department of Energy
Lawrence Livermore National Laboratory

DISCLAIMER

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Lawrence Livermore National Laboratory
Integrated Computing and Communications Department

FY02 I/O Integration Blueprint

Public Release Version 1.1

November 16, 2001

Kim Cupps
Mark Gary
Keith Fitzgerald
Terri Quinn
Dave Wiltzius
Kim Minuzzo
Mark Seager
Tyce McLarty

Revision History

Date	Comments
11/15/01	**Version 1.0 – K. Cupps** (Initial Draft)
11/16/01	Version 1.1 – K.Cupps incorporate Mark G.'s changes and add media to procurements.
11/27/01	Version 1.1 for Public Release – K. Cupps removed Appendix B information

This work was performed under the auspices of the University of California/Lawrence Livermore National Laboratory at LLNL under contract no. W-7405-Eng-48.

Table of Contents

I/O Integration – Executive Summary

This document describes I/O focused requirements, issues, options, plans, deliverables and budgets for Livermore Computing (LC) in FY02. Areas covered include I/O for archival storage, network, platform, visualization and the I/O Testbed. Implementation Plan (IP) milestones and tasks in each of these areas map to the efforts and plans described in this document.

When developing FY02 I/O requirements, a survey of key LC customers was performed (see Appendix A and D) and DisCom2 requirements were gathered. The LC customer provided throughput and capacity estimates were quite conservative when compared to ASCI curve projections and were history-based rather than being based on hardware capabilities. Because substantial differences exist in the ASCI platform in FY02, required I/O throughput rates were raised appropriately (i.e., by over 200% platform-to-archive). Archive capacity requirements remain fairly stable in FY02 as aggressive FY01 plans and purchases will accommodate most of the volume of data received through FY02.

10 Gigabit Ethernet network infrastructure will begin to get deployed in early FY02. When full implementation becomes cost effective it will allow us to greatly increase bandwidth between computer facilities. In addition pre-production OC-48 Ultrafastlane encryptors will be installed in early FY02 at each of the Tri-Labs. Preliminary testing of these encryption units and production unit installation will also take place in FY02. Significant network tuning, routing and protocol efforts will be investigated during FY02. These include channel bonding, gigabit ethernet flow control and HPSS protocol enhancements.

Changes to and investigation of archive mover platforms and tape technologies are planned in FY02. IBM tape robotics and tape devices will be phased out. New Storage Area Network (SAN) and platform-based mover architectures will be prototyped and investigated as approaches capable of significantly reducing archive expenditures in support of future 60 or 100 TFLOP platforms. Disaster recovery and RAIT device strategies will be formulated and HTAR porting (machine-to-machine and Linux) and deployment efforts will be completed.

The I/O Testbed will continue to be instrumental in pursuit of procuring reliable high performance hardware and software solutions. The Testbed will be used to provide a proof-of-concept for both the direct-to-tape and HPSS SAN investigations, as well as for benchmarking new node and disk platforms and storage devices.

Implementation of the FY02 I/O Blueprint will require $2.2M in network investments, $6.25M in archive improvements and a $.37M I/O Testbed investment. Manpower costs will remain constant. The requirements justifying these costs are found in Appendix A. The procurements are detailed in Appendix B.

Purpose

This document defines the necessary architectures and implementation plans for achieving required end-to-end I/O services between the ASCI platforms (White, SKY, EDTV), the archive storage system (HPSS), SMP servers (Compaq cluster), visualization servers (whitecap, tidalwave) and remote tri-Lab systems within both the Secure Computing Facility (SCF) and the Open Computing Facility (OCF), as specified in the "FY02 I/O Integration Transfer Rates and Capacity Requirements" in Appendix A.

Scope

This document specifies the technical attributes or I/O services to be provided in the SCF and the OCF including:

1. Aggregate and point-to-point transfer rate requirements between the five major end nodes: ASCI compute platforms, data archive (HPSS), SMP compute platform, visualization servers, and tri-Lab users. These requirements are fully described in Appendix A.

2. Network topology and technologies, including end-node network configurations.

3. An analysis of I/O issues and recommendations for addressing these issues in the areas of networks, storage and visualization.

4. Procurements are contained in Appendix B, and tasks for implementation are contained in Appendix C.

Primary areas of emphasis for I/O integration in FY02 are:

1. The enhanced I/O capabilities to be deployed in SCF required to support 25 TeraOPs of ASCI platforms.

2. Prototyping strategies for the reduction of future archive hardware expenditures.

3. I/O Testbed studies of SAN hardware and software approaches, node, disk and tape hardware.

4. DisCom2 WAN performance improvements and new technology/architecture deployment.

Level One Milestones

A primary goal of this Blueprint is to document the I/O architecture necessary to support "Level One Milestones". The Level One milestones important to the LLNL Director and Associate Directors are high-level milestones that will, in part, determine whether the Laboratory succeeds in fulfilling its major missions. Milestones at this level have been separated into "terascale science" and "production services." Terascale science activities push the envelope of computer science and performance requirements. Production services activities are needed to provide the operational computer center support to enable terascale science calculations.

Terascale Science Milestones

1. NA-2.1 FY2002 Q1 Three-dimensional prototype full-system coupled simulation [Nuclear Applications]

2. NS-2.1 FY2002 Q4 Three-dimensional safety simulation of a complex abnormal explosive-initiation scenario [Nuclear Safety]

3. NN-2.1 FY2002 Q4 STS abnormal environment prototype simulation [Nonnuclear Applications]

4. PM-2.2 FY2002 Q4 Delivery of initial macro-scale reactive flow model for high-explosive detonation derived from grain-scale dynamics [Materials and Physics Modeling]

Production Services Milestones

1. None for ASCI

Platforms Milestones

1. PP-2.1 FY2002 Q3 30-teraOPS system (Q), final delivery and checkout

The projected I/O requirements specified in this document are designed to achieve the Level One milestones listed above. Please refer to the "FY02 I/O Integration Transfer Rates and Capacity Requirements" in Appendix A for a full description of the I/O requirements. The FY02 I/O specifications that will be achieved by the end of FY02 are summarized in Figures 1 and 2 and Tables 1–3 below. Note that in some instances these specifications are lower than those stated in Appendix A. The overall strategy is to make the procurements necessary to achieve the requirements as specified in Appendix A; however, in some cases these requirements exceed what we can afford to provide.

The specifications given in this document represent the "guaranteed-to-be-achieved" numbers by the end of FY02, some of these differ from the requirements. It is important to understand that these rates are aggregate throughput rates and will not be seen on single file transfers. Several concurrent sessions running on several nodes may be required to achieve these aggregate transfer rates.

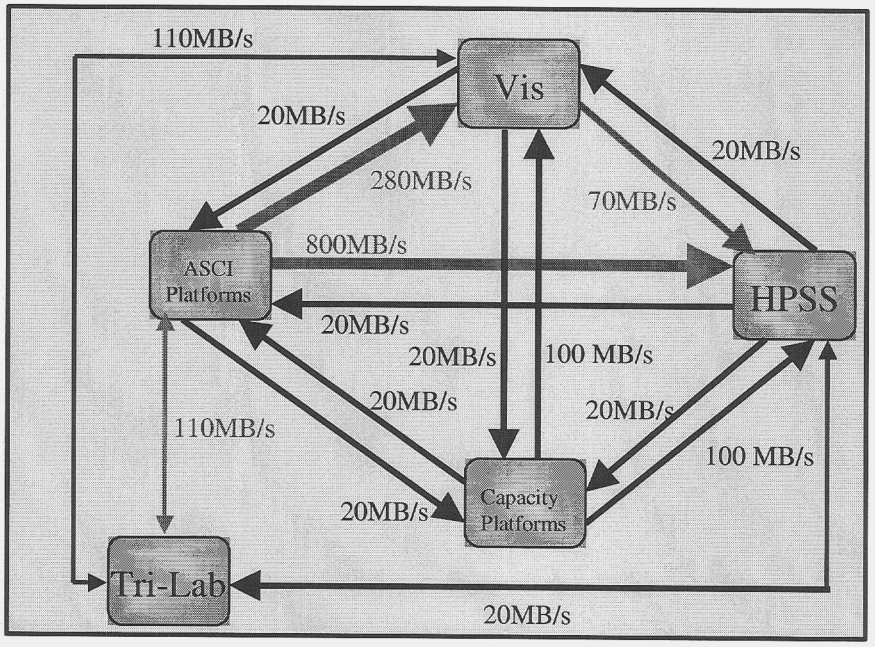

Figure 1. SCF Throughput Specifications

Note: The red lines in Figures 1 and 2 are for those connections considered to be of primary importance.

An extensive amount of tuning and software development is necessary to achieve the required end-to-end performance on all the links. Given the large number of variables and unknowns, the tuning must basically be accomplished empirically, thus requiring a significant amount of time.

The SCF DisCom WAN is capable of a maximum 240 MB/s bidirectional throughput. Figure 1 shows that this capability has been spread between three links to the WAN. The OCF DisCom WAN is capable of a maximum 12 MB/s bidirectional throughput, again Figure 2 shows the specifications distributed among three links to the WAN. The throughput was distributed over the three connecting links somewhat arbitrarily. The DisCom model is that a user is entitled to the entire bandwidth of the WAN on any particular transfer, however there is no coordinated scheduling of transfers from separate platfroms over the WAN.

SCF Aggregate Transfer Rate Specifications (MB/s)			
System	Write	Read	Total
ASCI	1210	170	1380
Vis	220	510	730
HPSS	80	990	1070
Compaqs	120	60	180

Table 1. FY02 SCF end-node aggregate transfer rate specifications.

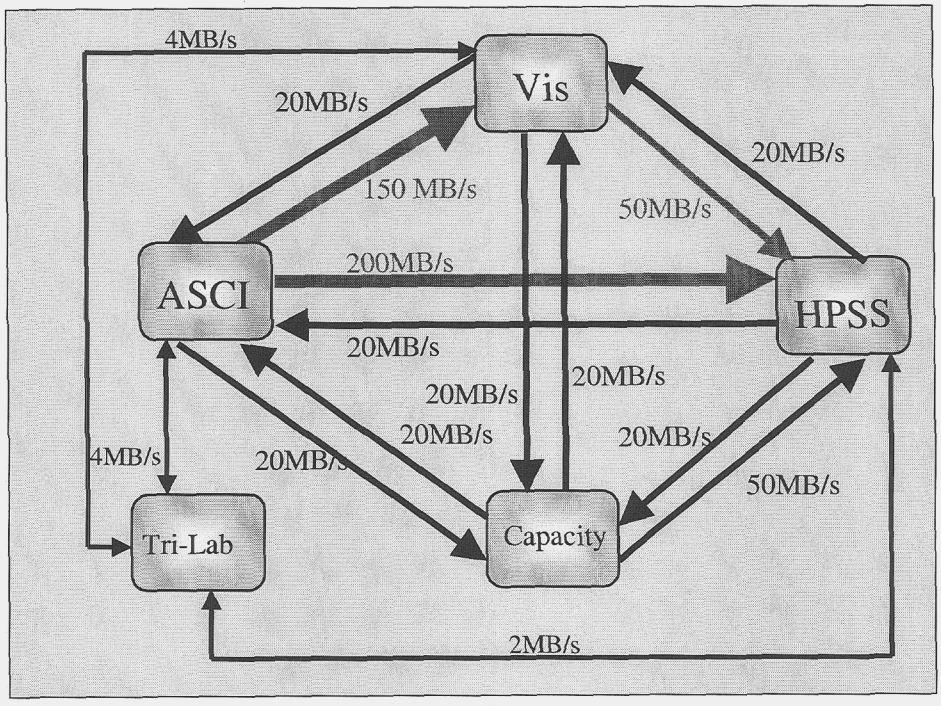

Figure 2. OCF Throughput Specifications

SCF Aggregate Transfer Rate Specifications (MB/s)			
System	Write	Read	Total
ASCI	374	64	438
Vis	94	174	268
HPSS	62	302	364
Capacity	90	60	150

Table 2. FY02 OCF end-node aggregate transfer rate specifications.

FY02 Issues, Analysis and Recommendations

Archival Storage Issues

This section discusses archival storage issues to be addressed in FY02. Each issue is described and then recommended actions are presented.

Issue A: High Archive Costs to Support Future Platforms

As we contemplate how to architect an archive that is capable of providing reasonable throughput of data generated by machines of up to 100TFLOPs of capability, it becomes obvious that our current model is very expensive when we scale even to 60TFLOPs. Further, we know that an ideal environment would have a Scalable Global Parallel File System capable of managing all of the disk on a network. The current vision for the next generation (2-3 years out) ASCI system at Livermore is many platforms and devices connected to an Infiniband SAN, magically moving data when and where it's requested. We will put a significant amount of effort this year into investigation of HPSS alternatives that move us closer to this SAN vision. Figure 3 depicts our current vision for the next generation ASCI system at Livermore.

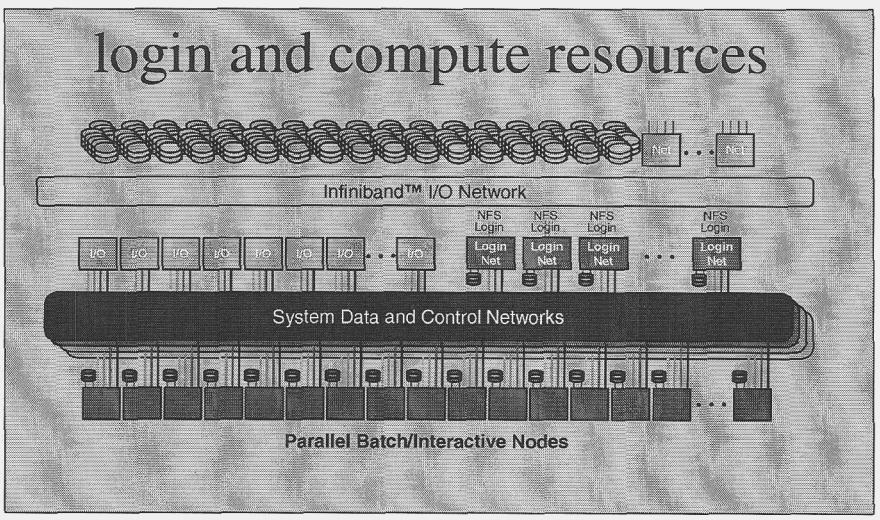

Figure 3: Projected ASCI system architecture

FY02 Action/Recommendation

We plan to investigate several alternatives aimed at reducing outyear archive costs as well as advancing us toward the SAN vision of the future.

- **Improved IBM Mover Platform** - This entails borrowing a late generation IBM node and doing performance tests aimed at determining whether we can use these nodes as data movers and reduce our overall costs. The deliverable out of this exercise is a write-up of the test plan results and a determination of the new mover platform's capability as compared to an IBM SP Winterhawk, currently our most powerful archive data mover platform. This study will directly influence FY02 archive platform purchases.

- **HPSS SAN Strategy Investigation –** This is close to a yearlong effort and requires modification to local HPSS source code as well as modification to vendor device code. The basic steps are outlined below and a whitepaper outlining an evolutionary strategy is available. This plan would eventually lead to the elimination of movers from the data path by using the third-party copy protocol and device-to-device copies. If this implementation is adopted it will reduce the number of mover nodes required to transfer data to archive disk and tape devices resulting in a significant outyear cost reduction. FY02 tasks include:

  - Prototype the elimination of the disk mover platform from the data path for tape migration.

  - Complete cost/benefit analysis of FibreChannel storage directors as a front-end for disk and tape devices and purchase/deploy based on study results. These directors promise to improve availability, scalability, connectivity, management flexibility and represent a step toward deploying a SAN environment for the archive. Unfortunately these directors are expensive.

  The deliverable from this investigation is a write-up of the results of the testing including a quantification of mover node reduction and cost reduction factors.

- **Investigate mover running on ASCI platform node –** This investigation will center on putting movers directly on the ASCI platform and using the SAN fabric to create a very wide stripe to tape while identifying performance and reliability issues. Wide stripes to tape are necessary in order to get reasonable performance on large files. Mirroring would be necessary in a production implementation unless RAIT was available. Without RAIT capability, tape reliability is a major factor when considering how wide to stripe.

  HTAR is another issue that comes up when considering this implementation. Currently HTAR cannot be used to write directly to tape. The HTAR developer has indicated that this is an extremely complex issue and one not easily solved. This means small file writes with this implementation will be costly. If this strategy is proven to be efficient and reliable, it would result in large outyear cost savings of both disk and disk mover nodes. The basic steps involved in implementation of this investigation are outlined below:

  - Develop test plan.

  - Install and test the FTP code capable of invoking the HPSS Local File Mover (LFM) on a testbed compute platform node

  - Install and test the HPSS Local File Mover (LFM) code on a testbed compute platform node

  - Install and test a SAN switch connected to a host bus adapter on the testbed compute platform and connected to tape devices in the Testbed's silo.

  - Execute the test plan for this implementation.

  - Deliver a written cost/benefit analysis of this approach and make a recommendation

- **Examine low cost LINUX mover platform alternative –** We will investigate whether moving to a commodity mover platform is a viable strategy to reduce outyear equipment costs.

- Develop test plan.
- Put an HPSS mover on an Intel LINUX box.
- Execute the test plan for this implementation. Much of what was done with the IBM mover platform can be reused here.
- Deliver a written cost/benefit analysis of this approach and make a recommendation

Issue B: Small File Sizes

As disucssed in detail in the FY01 I/O Blueprint, writing thousands of small files to HPSS poses significant throughput issues. Multiple and lengthy FTP sessions are required and are time consuming for users. Additionally, HPSS write performance is dominated by the file create time when writing many small files. Small files are still being generated due to the same reasons outlined last year: continued use of Silo and delayed incorporation of DMF into user's codes. The HPSS archive statistics show that small files still account for 90% of the file transfers to both the OCF and SCF archives, but account for less than 10% of data transferred. Figure 4 shows OCF file transfer statistics by file size for a typical week in 2001.

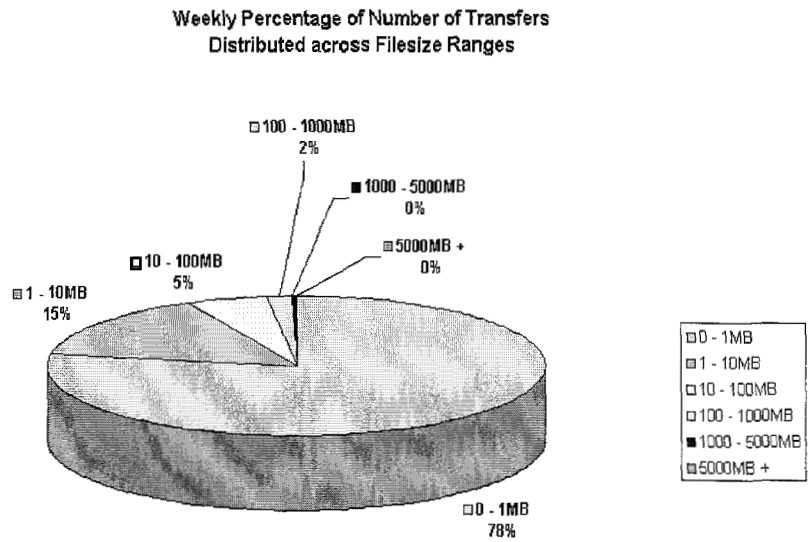

Figure 4. OCF File Transfers by File Size 10/7/01

There was considerable headway made on this issue in FY01 (a factor of 46 small file throughput improvement), with deployment of the HPSS Hierarchical Tape Archiver (HTAR), and re-hosting of the HPSS Meta-data servers on faster hardware. However, many users are still transferring thousands of

small files to the archive without using HTAR. Our new challenge is to get the majority of our users to adopt HTAR and use it regularly.

FY02 Action/Recommendation

Target users for adoption of HTAR: We must get more users to use HTAR. To do this we will assign one person in the Data Storage Group to be the HTAR user liaison. This person will be responsible for training users to effectively make use of HTAR. We plan to use our transfer logs to identify ten key people to target as HTAR adopters. This is more difficult than it may sound as our users are busy and like most people, like what they know. HTAR's incredible performance will make the pain of learning a new interface less than the pain of long transfer times.

Other HTAR tasks that must be completed this year include porting HTAR to LINUX so that it is available on the new LINUX clusters, and creating a machine-to-machine version of HTAR. The machine-to-machine version of HTAR will be used to tranfer files between non-HPSS platforms.

Issue C: Lack of RAIT Capability

Redundant Arrays of Inexpensive Tape (RAIT) are vital to our ability to mitigate tape damage risk. It is also crucial to our ability to send large files to widely striped tape cost effectively (without duplicate copies) and reduce our dependence on disk as a speed-matching buffer. The StorageTek RAIT project has not met either its deadlines or its performance objectives. Creation of parity in hardware and transfer of that parity data to tape has proven to be a tough problem for StorageTek. The current schedule calls for StorageTek to deliver RAIT parity generation at 80MB/s via the SN6000 product by September of 2002. This is a year after it was originally promised and there is little encouraging evidence the schedule will be met.

FY02 Action/Recommendation

Seek a RAIT tape solution. Continue to follow/encourage the StorageTek RAIT project and investigate alternative RAIT solutions. Follow the progress of StorageTek's virtual tape mirroring box. If the product delivers in a timely fashion investigate the cost-effectiveness of either deploying the device as a direct-to-tape device for very large files and/or placing mirroring devices behind a disk cache for very large files.

Issue D: Disaster Recovery

Recent world events have focused attention on the need for providing geographically distributed copies of critical archive data. While the current LLNL HPSS systems keep multiple copies of many files, these copies are stored in close proximity to each other. Which files get multiple copies is currently determined based on file size and user interface.

Because of the large volume of data ingested daily into the HPSS archive any proposed solution would have to involve explicit selection of which data is critical in order for the implementation to be affordable. In addition, because of the potential scale of disaster scenarios, second (backup) copies of critical files should not be housed physically at LLNL. Los Alamos National Laboratory (LANL) has these same concerns and is interested in mechanisms for storing critical data at LLNL or at another remote/secure location.

FY02 Action/Recommendation:

In FY02 a detailed analysis of the feasibility (operational, user, cost) of remote, secure storage of critical data will be completed. A number of mechanisms have been suggested for accomplishing this:

1) WAN transfers of copies of critical data to/from LANL.
2) Shelf storage of copied tapes at the Nevada Test Site (NTS).
3) Shipment of second copies of critical data to/from LANL.
4) Siting of remote data movers at LANL.

Many technical, operational and budgetary challenges exist with each scenario. Common to all options is the need to identify which data is critical or essential. The analysis performed will generate a brief report detailing recommended actions.

Issue E: High-Speed Archive Throughput

The continued presence of the ASCI SKY and White Machines and the addition of EDTV and Linux clusters yield daunting FY02 archive throughput requirements. The files sent to HPSS must be transferred at high rates in order to avoid resource contention on the capability and capacity platforms. The fact that the fastest archival storage devices store data at the lowest density and greatest cost demands that a heterogeneous mix of devices and hierarchies be maintained.

Given the FY02 throughput and capacity requirements, the present archive implementations will be inadequate because:

- Existing disk caches are not large enough and don't provide enough throughput to buffer large transfer loads from ever-expanding platform disk and I/O capabilities.

- With the advent of files greater than 100GB the amount of time a file takes to transfer to two-wide striped tape devices is becoming too great (>4 hours). This increases the time that large files occupy the disk cache and creates extremely large stage times. Unfortunately it has become clear that the StorageTek RAIT project is behind schedule and will only be able to offer a mirroring implementation during FY02. In addition fast optical tape technology (LOTS) continues to be at least 18 months away. The combination of very large files and stagnant tape transfer rates presents the archive with serious challenges.

Action/Recommendation

In FY02 we propose the following changes to the archive hardware architecture:

1) Expand the size and bandwidth of the HPSS disk cache in the OCF and SCF. Presenting the highest throughput devices as the top level of the hierarchy will continue to be our goal this year. Because platform disk caches continue to grow, and because disk prices continue to drop dramatically and RAIT and SAN technology are not quite here, this approach is well supported by technical and financial arguments.

2) Continue to deploy StorageTek 9940 tape technology behind a majority of the HPSS disk cache. Surveys of competing technologies show the 9940 to be the best choice for our archive. As soon as 9940B (200GB at 30MB/s – see Table 3) technology becomes available procure these drives and exercise options to upgrade as allowed by procurement restrictions.

3) Phase out IBM 3494 Robot and IBM 3590E technology. It has become clear that our present IBM 3494 robots do not present upgrade paths for even IBM's future tape technologies. The robots and drives will be phased out of production.

4) Expand the tape stripe-width behind HTAR. Files greater than 100GB in size are taking much too long to migrate to tape and will take a commensurately long time to be staged. Without faster tape technologies, and with the failure of the RAIT program to deliver, we will need to increase the width of our tape stripes and incur the added reliability risk that this entails.

Tape Type	Native Capacity	Compressed Capacity	Transfer Rate	Availability
IBM 3590E	20 GB	30 GB	15 MB/S	Now
STK 9840	20 GB	30 GB	10 MB/S	Now
STK 9840B	20 GB	30 GB	20 MB/S	Late '01
STK 9940	60 GB	90 GB	10 MB/S	Now
STK 9940B	200 GB	300 GB	30 MB/S	First Half '02

Table 3: Tape Technology Overview

We believe that implementation of the FY02 action and recommendation plans outlined above will provide the highest possible data rates to our users while minimizing media cost and footprint by leveraging emerging tape products. Several of these strategies represent the first steps toward a true archive SAN implementation in preparation for future platform deployments. Testbed efforts evaluating processor, RAIT and other tape technologies, combined with close attention to developing technology (SAN and tape) will position us well for FY03 and beyond.

DisCom2 Issues

DisCom2 tri-Lab services for SecureNet and the newly deployed OC-48 DisCom2 WAN (2.4Gb/s) each present their own challenges, issues, and requirements.

Issue F: Grid Services (formerly DRM)

The DisCom2 program supports the development and, as appropriate, the deployment of Grid Services. The low-level grid service is provided by Globus. The grid services will need to communicate between the tri-Lab sites (and NWC plants). This inter-site communication will by affected by firewalls. The DisCom2 Grid Service project may suggest some strategies to minimize the impact of firewalls on the deployment and performance of the Grid Services. These strategies need to be coordinated with the software and hardware architecture deployment.

FY02 Action/Recommendation

In FY01 LLNL deployed a Grid Services server for White. We expect to expand this service to include SKY. Further, LLNL expects to deploy a Grid Services client and hope to get some LLNL users to evaluate the Grid Service for doing production work on the Q ID system ("QSC"). These services will be on the classified network which does not have firewalls that are troublesome for Grid Services.

Issue G: Tri-Lab Data Movement

The tri-Labs have identified a standard FTP client that can use the DCE certificate for authorization (as GSSFTP), and is able to do multiple parallel streams to transfer a file (as PFTP). In FY01 the DisCom2 program coordinated the deployment of the GSS-PFTP client and server to most tri-Lab ASCI computing resources. Furthermore, to meet the DisCom2 FY01 milepost it was required that HPSS make changes to the HPSS mover protocol for the HPSS FTP server.

In response to the architecture of the Q machine LANL is embarking on a multi-year collaborative effort with ANL and the tri-Labs to develop and deploy "GridFTP." The proposed GridFTP clients and services are expected to support DCE credentials and parallel paths, but will additionally support the concept of multi-node PFTP. The multi-node PFTP capability will likely utilize MPI-IO to accomplish inter-nodal data movement. The other tri-Lab facilities are expected to deploy a GridFTP client, server or both to provide high-performance access over the DisCom2 WAN to LANL's Q machine.

FY02 Action/Recommendation

Few developments are expected for the existing GSS-PFTP client and server beyond what has already been deployed. However, we do hope that HPSS will incorporate many of the features into the HPSS FTP client and server, particularly the HPSS modifications to the mover protocol.

Additionally, LLNL will participate in the development of the GridFTP client and server. This will probably be a multi-year effort.

Issue H: Remote Visualization

New DisCom2 user requirements emerged with the allocation of 32 nodes on White for VIEWS applications. LANL and SNL users of Ensight are utilizing the VIEWS nodes to partially render images on White and transfer the Ensight geometry data file to a local SGI for rendering at the user's desktop. It is important that this usage model be fast.

FY02 Action/Recommendation

An important first step was taken in July 2001 by installing a jumbo frame Gigabit Ethernet interface on each of the VIEWS nodes. As more users are allowed on White, and users become more familiar with the DisCom2 WAN, we expect the network connection of the VIEWS nodes to generate considerable traffic. Still, we have been told that the performance experienced by LANL Ensight users on White is comparable to what they experience with all elements of Ensight running locally on their SGIs. So the greatest throughput increase from the VIEWS nodes may be when more users have access to White and the VIEWS nodes.

The tri-Lab visualization resource Whitecap will be brought online in early FY02. It is unknown how much this resource will be used and the type of DisCom2 WAN traffic generated by it.

Issue I: DisCom2 WAN Usage and Usage Statistics

The DisCom2 program, LLNL included, is gathering statistics on the usage of the DisCom2 WAN. These statistics indicate that 4-5 days a week the DisCom2 WAN is "heavily" used (e.g., 100MB/s) for a few hours of the day. For example, in August about 12TB of data was transferred over the DisCom2 WAN (even though the network was idle for roughly 50% of the time). Unfortunately the statistics gathered are not adequate.

FY02 Action/Recommendation

The gathering and presentation of usage statistics at LLNL will be made more thorough. The goal will be to understand the traffic for each type of source (e.g., VIEWS nodes, White login nodes, SKY, etc) and, to the resolution of which site, the sink. Since we can only gather information from our local network components, we do not expect in FY02 to gather statistics on even the layer 2 information (adapter source and sink), but we will investigate some possibilities. Still, this will help us understand the traffic patterns in LLNL and to/from the DisCom2 WAN at LLNL.

Issue J: WAN Architecture Limitations

Currently, heavy usage of the tri-Lab WAN is limited by the sinks at the other Labs. This represents a rather simplistic use of the WAN by users that almost fully utilize White. Further, hardware available for the WAN network components is limiting per machine throughput since only standard Ethernet packets are supported.

FY02 Action/Recommendation

Each Lab is working hard to upgrade their network and computer infrastructure to better utilize the DisCom2 WAN. For example, LANL expects to upgrade their HPSS mover machines and use Gigabit Ethernet in early FY02. Each Lab also expects to upgrade the WAN network hardware to support jumbo frames in early FY02. In early FY02 the tri-Labs expect to upgrade the WAN network hardware to support jumbo frames, and the per machine/node sustainable throughput should increase from 100MB/s to about 250MB/s (very close to the 270MB/s theoretical maximum for the OC-48 (2.4Gb/s) DisCom2 WAN).

In FY03 DisCom2 expects to upgrade the WAN capacity from one OC-48 to two OC-48 (about 540MB/s capacity). In that timeframe Q should be deployed, as should EDTV.

Off-Platform Visualization I/O Issues

This section discusses off-platform visualization I/O issues to be addressed in FY02. Each issue is described and then recommended actions are discussed.

Issue K: Platform – Vis File Transfers

File movement between platforms and the visualization server is important to users, since the visualization servers are used for analysis of results. Our requirements, as identified in the FY01 I/O Integration Transfer Rates & Capacity Requirements were based on ASCI White simulation and as such remain unchanged for FY02. The requirement is to move 6TBs of data from White to the classified SGI server in 6 hours resulting in a 280MB/sec transfer rate and to move a 6TB data set from the SGIs to HPSS in 1 day resulting in a 70MB/sec transfer rate.

We have installed the hardware and parallel FTP (e.g., GSS-PFTP) clients and servers to support these requirements. Each SGI has 4 jumbo frame Gigabit Ethernet interfaces and 2 Gigabit Ethernet interfaces for NFS traffic (e.g., internal) and for common user access (e.g., external) characterized by SSH, serial FTP, X-windows, etc. With this architecture and software we have demonstrated over 200MB/s from White to an SGI visualization server (e.g., Tidalwave) over the 4 jumbo frame Gigabit Ethernet interfaces.

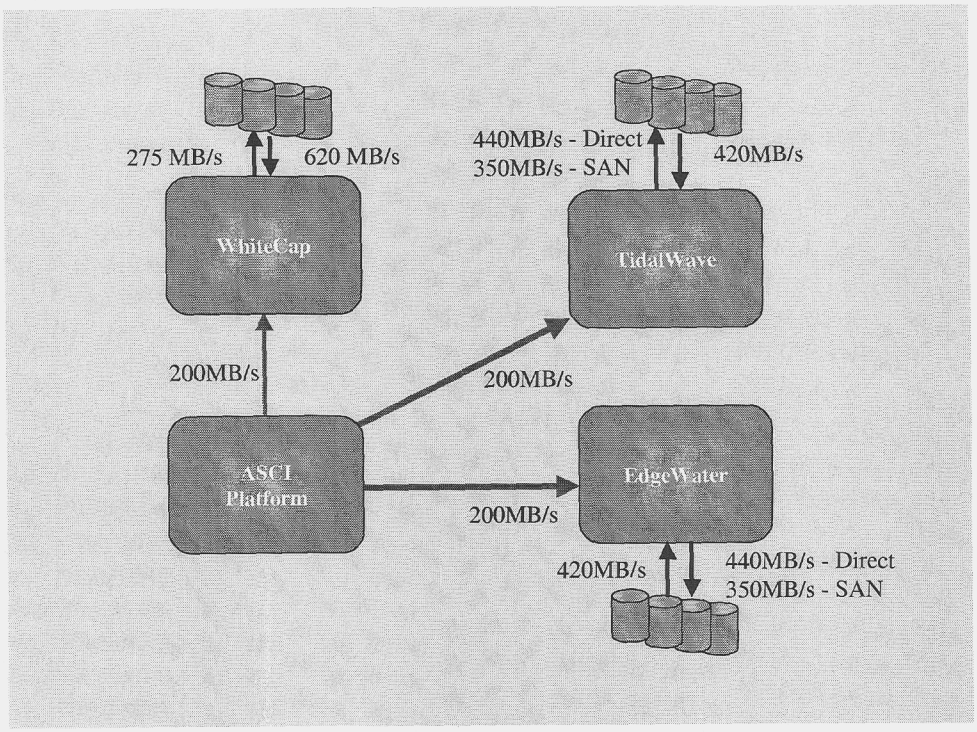

Figure 1: FY02 ASCI Platform to SGI Visualization aggregate transfer rate specifications

Action/Recommendation

It is recommended that machine-to-machine HTAR be developed to support the transfer of groups of smaller files from one machine to another, specifically from White to visualization servers. It is also recommended that performance numbers for NFT between White and the Viz are quantified and published. Table 4 highlights platform-to-vis transfer issues.

Utility	File Size	Numbers of Files	Status
FTP	< 1GB	Few	Deployed but needs tuning
NFT or Htar	< 1GB	>100	NFT deployed , but no performance numbers Machine-to-machine HTAR does not exist
PFTP	> 1GB	N/A	Performance may be improved if optimized for SGIs.

Table 4: Software Needed for Visualization I/O

To address these issues in FY02 we plan to take the following actions:

- Continue to tune FTP
- Test NFT and HTAR to the SGIs and publish performance numbers
- Develop and deploy machine-to-machine HTAR

- Periodically monitor the performance of the data transfer utilities to and from the SGIs

Issue L: Network Performance, Stability and Configuration

As bandwidth requirements continue to grow year-over-year network performance and stability become increasingly important. Although we have made significant increases in these areas recently we will continue to investigate strategies to improve performance, stability, and configuration.

Action/Recommendation

1) **Channel bonding**. Channel bonding has the potential to aggregate 2 or more network interfaces as one virtual IP interface with increased capacity (e.g., 2 Gigabit Ethernet interfaces bonded as one IP address with 2 Gb/s capacity). This technology will be investigated in FY02 and a short report on findings will be delivered.
2) **Gigabit Ethernet flow control**. Some interpretations of this feature imply that no packets would be lost in the Gigabit Ethernet paths found at the SCF and OCF. This feature should be understood and utilized to the fullest. This technology will be investigated in FY02 and a short report on findings will be delivered.
3) **10 Gigabit Ethernet.** Plans are to deploy 10 Gigabit Ethernet by Q3FY02 as a pair of trunks between the B113 and B451 areas on SCF. When cost effective 10 Gigabit Ethernet will allow us to greatly increase the bandwidth between computer facilities anywhere on campus.
4) **Network WAN hardware.** We've encountered several problems in the Cisco 8540 network WAN router:
    a) Working with the hardware vendor (Cisco) modifications were made to IOS to allow "bonding" of multiple OC-12 interfaces so that the realized capacity of the 4xOC-12 interfaces would be a bit over 2 Gb/s using only 2 Gigabit Ethernet ports. We are now awaiting this feature in a production version of the IOS. This will not change the expected capacity of 270MB/s of the DisCom2 WAN, but will decrease the number of Gigabit Ethernet interfaces required in the network hardware to achieve this. This is important when we make the next step to 2xOC-48 capacity for the DisCom2 WAN as it saves considerable costs for WAN hardware for that step up in throughput.
    b) In FY02 we will be testing the OC-48 Ultrafastlane encryptors in preparation for deployment in late FY02 or early FY03. Currently the problem resolved above for OC-12 ATM interfaces still exists for OC-48 ATM interfaces. We'll be working with Cisco to ensure the scheduler and bonding solutions available for OC-12 interfaces will become available for OC-48 interfaces.
5) **OC-48 Ultrafastlanes.** Pre-production units of the OC-48 Ultrafastlanes will be available in early FY02, with one going to each Lab. Preliminary testing of these encryption units will begin at that time. Additional, production units are expected in Q3FY02. These will be used to test the Ultrafastlanes and WAN network hardware for 2xOC-48 operation in FY03.

All of the issues above have been considered for FY02, each is addressed in the schedule of work contained in Appendix D. These issues do not necessitate fundamental archive or network architectural changes. FY02 architectures are described in the following section.

FY02 Network Architecture

Figure 6 shows the current SCF hardware and network architecture. Large Gigabit Ethernet switches have been deployed to cost-effectively provide multiple, non-blocking, high-speed paths between computing resources and storage. The connectivity between platforms and storage is provided by two 176 port Cisco Catalyst 6513 Gigabit Ethernet switches (one in B113 and one in B451). A similar architecture and network design is used in OCF.

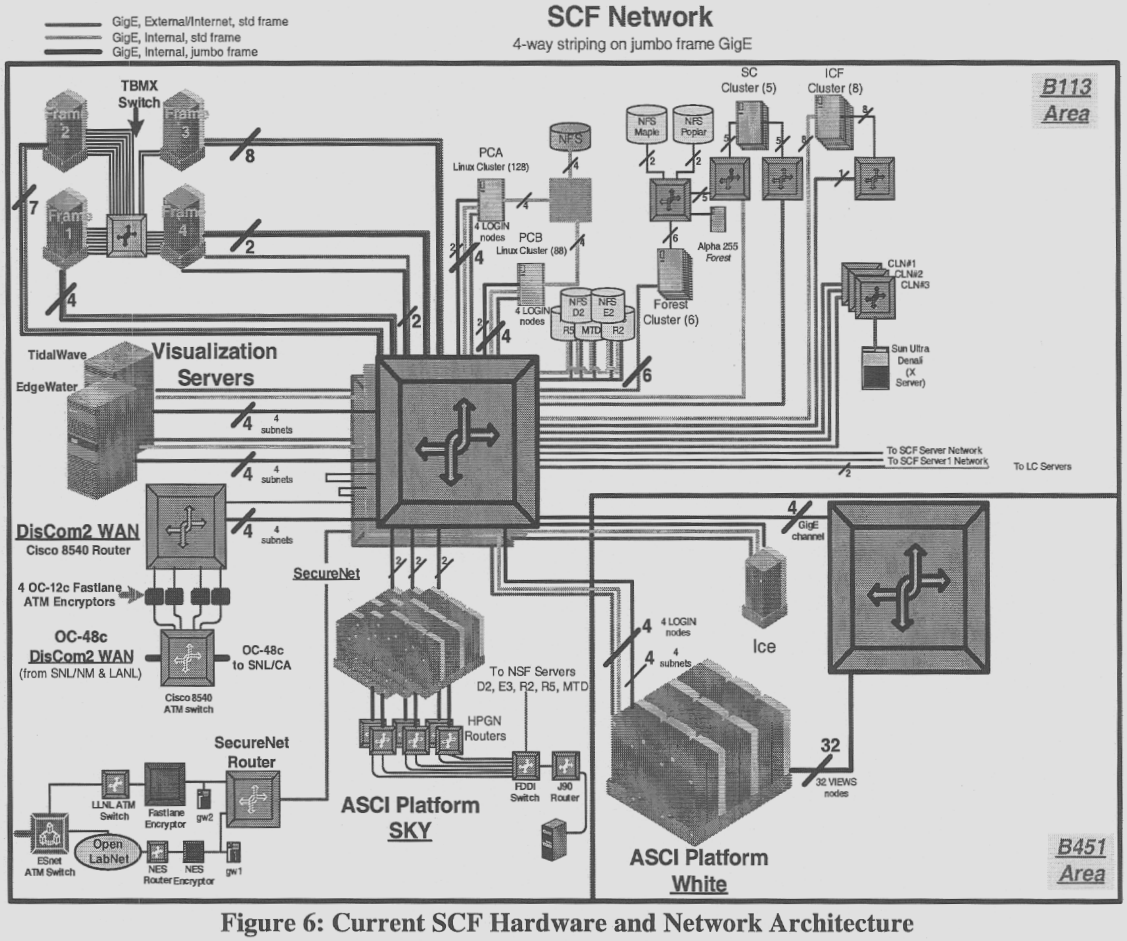

Figure 6: Current SCF Hardware and Network Architecture

HPSS Architecture

The basic node and network architecture of HPSS remains the same as that specified in the FY01 I/O Blueprint. Figure 8 shows this architecture with disk capacities that will be achieved by the end of FY02. Tables 5 and 6 show the current and target FY02 HPSS infrastructure. The costs associated with providing this increased infrastructure are outlined in Appendix B: Procurements.

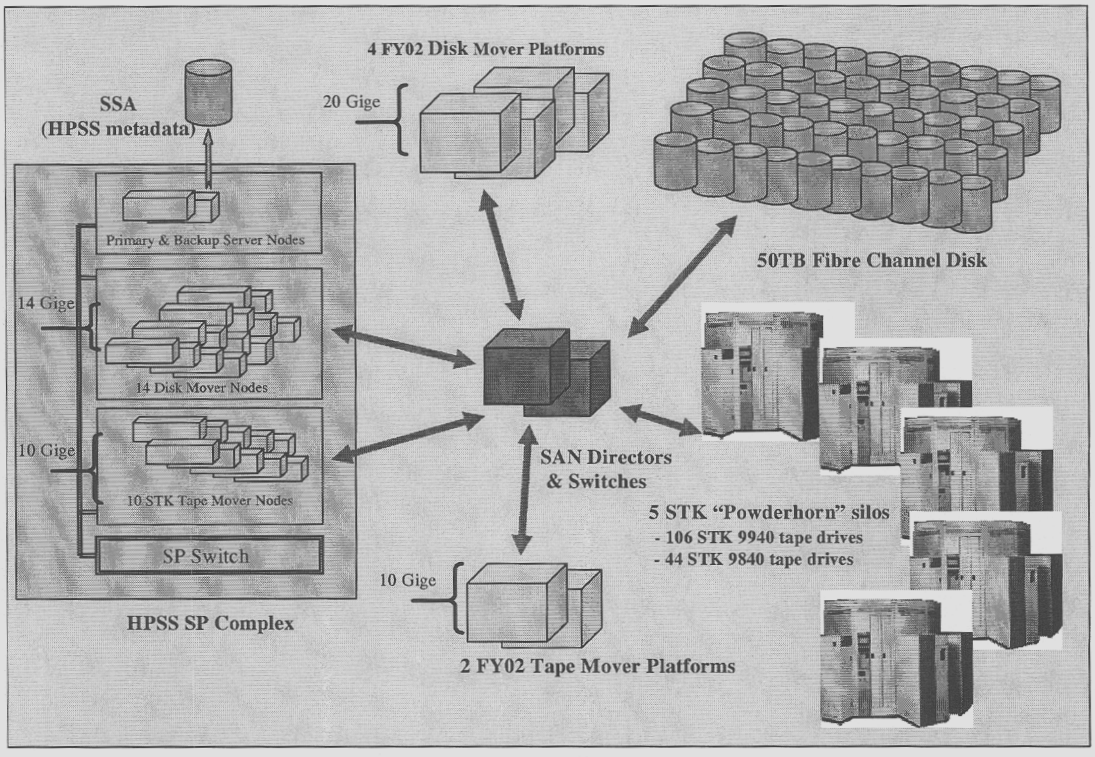

Figure 7: High Level Architecture of FY02 SCF HPSS

Parameter	FY01	FY02
Slot Capacity	2.7 PB	2.7 PB
Tape Capacity	969PB	1.6 PB
Disk Cache	40 TB	50 TB
Archive Nodes	32	38
Aggregate Peak BW (MB/s)	600	1430

Table 5. SCF Storage Current & Target Infrastructure

Parameter	FY01	FY02
Slot Capacity	1.36 PB	2.16PB
Tape Capacity	463 TB	563 TB
Disk Cache	25 TB	35 TB
Archive Nodes	24	27
Aggregate Peak BW (MB/s)	290	400

Table 6. OCF Storage Current & Target Infrastructure

The FY02 archive implementation is relatively stable when compared to FY01. The number of mover nodes, peripherals and size and connections to the HPSS disk cache will increase to provide required increases in throughput. Figure 8 describes the FY02 two-tiered tape architecture. The biggest change is the elimination of the 3590 tape and associated 3494 robots from our tape strategy. As discussed

previously, the IBM futures roadmap did not justify continued investment in the IBM tape direction. The StorageTek 9940B technology promises to provide a big win going from 10 MB/s per drive performance to 30 MB/s drive performance. The new drive will also increase cartridge capacity from 60 GB to 200 GB.

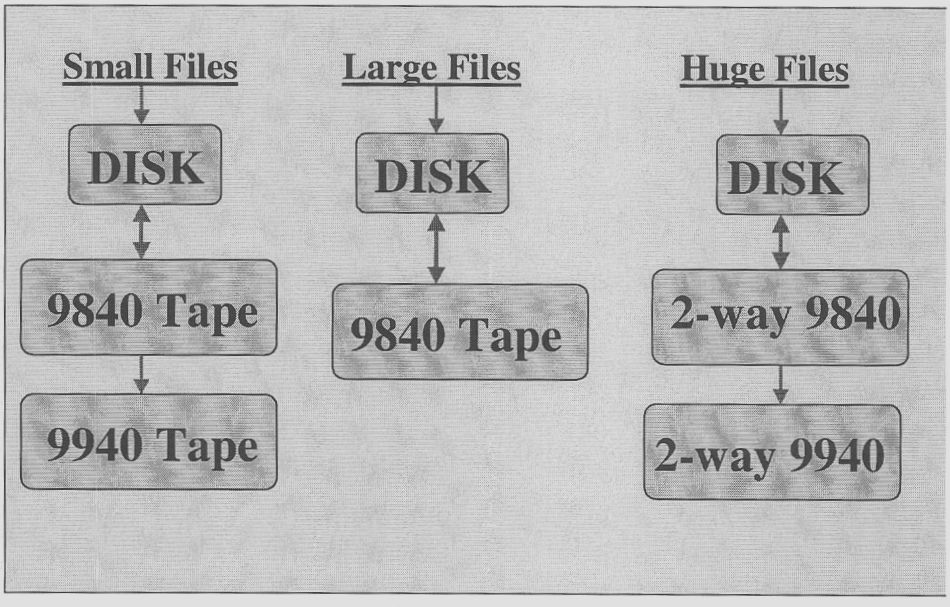

Figure 8. HPSS Storage Hierarchies

Appendix A: I/O Integration Requirements

This appendix describes the throughput requirements, both aggregate and point-to-point for the SCF and the OCF. The SCF will be made up of five major sub-systems connected by a network. These five sub-systems are the ASCI platforms, the archive, the capacity computing platforms, the visualization servers and the user workstations. The OCF also has five sub-systems that can be referred to in the same way, although the actual machines and their capabilities are not equivalent to those in the SCF. This document will also estimate archive capacity requirements in the SCF and the OCF. Capacity requirements will be estimated as a function of platform disk and memory sizes.

FY02 SCF Computing Platform Changes

The SCF computing platform will change significantly in April of 2002 with the addition of the Early Demonstration of Technology Vehicle (EDTV) system. EDTV will provide anywhere from 10 to 20 TFLOP/s of additional computing capability to the SCF, which means there will be between 75% and 125% additional ASCI platform computing power in FY02.

The FY02 Visualization platforms are Edgewater, Tidalwave and Whitecap. Whitecap is the Tri-Lab Visualization Server and will be newly operational in FY02. Whitecap doubles the compute capability of the visualization platforms, and almost triples the amount of memory. Whitecap has 6TB of disk cache, which ups the Visualization platforms total disk from 19TB to 25TB.

The capacity computing platforms consist of the ICE, Forest, Furnace and ICF cluster. These clusters are also undergoing a significant amount of change in FY02. The Forest Cluster is being upgraded. Four 8400s are being replaced by eight es45s. This is almost a wash in computing capability but more than doubles the memory of the Forest Cluster. A new, 213 GFLOP/s, loosely coupled Furnace cluster (64 node CS-20, Alpha EV68 running at 833 MHz with 100Base-T Ethernet interconnect) will come onto the floor near the end of CY01. The Parallel Capacity Resource (PCR) will also be deployed as two clusters into the SCF at the beginning of CY02. The tightly coupled P4A and P4B clusters will have 216 dual Pentium 4 1.7 GHz nodes and a peak of about 1.5 TFLOP/s, 432 GB of RDRAM memory, 17.8 TB local disk, 7.0 TB global NFS disk and Quadrics QsNet ELAN3 interconnect. Each of these clusters has 1GB of memory per processor and both will run Linux and OSCAR clustering software. The three new capacity computing platforms more than quadruple the previous compute power of FY01 SCF capacity platforms and provide a factor of 10 more memory. The PCR clusters also introduce a new "capacity" parallel job environment. It is anticipated that P4B will be migrated quickly to GA status on the SCF (hopefully by the end of October) and P4A will remain on the OCF side for three months doing science runs and then migrate to GA status on the SCF in January or February 2002.

Table 1 below lists the anticipated FY02 SCF machine characteristics of interest for the purposes of bandwidth and capacity requirements determination. The External I/O column is new this year and represents the aggregate off-machine transfer rate that can be sourced. It does not take into account any of the sink's abilities to accept data.

Machine	Compute Power	Memory Capacity	Disk Cache Bandwidth	Disk Cache Capacity	External I/O
White	12.3 TFLOP/s	8 TB	12.8 GB/s	120 TB	960 MB/s
SKY	3.9 TFLOP/s	2.6 TB	3.2 GB/s	75 TB	210 MB/s
EDTV	10 - 20 TFLOP/s	5-10 TB	12.8GB/s	120 TB	1.0 GB/s
Vis Platforms	76 GFLOP/s	148 GB	800 MB/s	25 TB	840 MB/s
PCR Clusters	1.5 TFLOP/s	432 GB	400 MB/s	7.0 TB	480 MB/s
Capacity Computing	992 GFLOP/s	320 GB	600 MB/s Local	4.5 TB Local	300 MB/s

Table 1: FY02 SCF platform configurations

FY02 SCF Computing Model

This section will consider four distinct SCF FY02 computing models. These models include the LLNL ASCI user computing model, the Tri-Lab ASCI user computing model, and the capacity computing model. The following assumptions concerning each computing model are key to the throughput requirements outlined in this document.

The local ASCI user's usage model described was derived from current usage patterns during Frost science runs, and projections from code developers about local use of LANL's Q machine when it becomes available. The local usage model follows. Large data sets are created on the ASCI computer and stored locally on the GPFS disk cache. We expect that a week long run will generate as much as 30TB of data to the GPFS disk cache. A user needs to get this data off of the GPFS disk in a "reasonable" amount of time. To get the data off of the machine in one-tenth the time it took to run the problem implies moving 30TB off of White and into HPSS in 16.8 hours. This assumes an I/O throughput rate of 495 MB/s. Most users do not archive all of the data they generate during a run. Assuming that a user stores only half the data that is generated, and this is a high estimate by historical standards, an I/O throughput rate of approximately 250 MB/s from the ASCI platform to the archive is required. Local users tend to keep their data on the GPFS disk cache as long as possible and do post-processing, including some software rendering on the ASCI computer. If allowed to complete post-processing on the ASCI computer, the data required to be shipped to the visualization servers is reduced. If it is required to move data off GPFS before completing post-processing the total data generated will typically be moved to the Visualization server for post-processing (hardware rendering and movie making), necessitating the full 495 MB/s throughput rate from the ASCI platform to the visualization server. The reduced results will then be stored into the archive from the visualization server. In either case, once the data is stored in the archive it is rarely retrieved.

Once Q becomes available, local users will compute there as much as possible. If VisIt is available, users plan to rely heavily on the client/server model - i.e. keep the data on Q, and display back to LLNL. The users with the special graphics hardware in their offices will probably try moving their data to TidalWave. LLNL users would prefer to archive the data they have computed on Q back in their home HPSS archive provided WAN transfer rates are not prohibitively slow.

The Tri-Lab usage model was derived from recent Crestone Project runs and discussions with Los Alamos National Laboratory (LANL) users. A Tri-Lab user logs into an ASCI machine using his foreign kerberos credentials. During the problem calculation large data sets are generated and stored on the local GPFS disk. Visualization of data is required as is tertiary storage of selected data sets. The visualization model has changed significantly since the acquisition of 32 nodes of White that are dedicated to visualization purposes. These are referred to as the "VIEWS nodes". The 32 VIEWS nodes represent a major reduction in visualization network throughput requirements. The totality of the visualization data is rarely (never) sent across the network to the Visualization server. Tri-Lab rendering is accomplished by running Multiple Ensight servers on the 32 VIEWS nodes of White to extract visible geometry data from the dump data on GPFS. The extracted visible geometry data is then sent over the WAN to a rendering client on a LANL visualization server. The visible geometry data is about 4% of the computational geometry data during a timestep, so the WAN is required to provide throughput for 4% of the total dump data generated.

The rendering model may change again with the addition of the Tri-Lab visualization server. LANL scientists are eager to use the new Tri-Lab visualization server. They plan to run rendering clients on the Tri-Lab server with the data provided by multiple Ensight servers running on the 32 VIEWS nodes. It is not clear whether this will increase or decrease WAN bandwidth usage. Even though images are much smaller than visible geometry, they will be sent more frequently.

The majority of the data generated during a Tri-Lab calculation is bundled into large tar files using HTAR and stored in the local archive. There has also been talk of creating an import/export capability

into HPSS to handle mass movement of data cartridges between Tri-Lab HPSS systems. This request is coming from LANL management and is thought to be a WAN risk mitigation strategy.

The capacity computing computational model has traditionally been quite different than that of both the ASCI and Visualization platforms. FY02 is the first year that there will be two parallel non-ASCI capacity computing platforms available. These parallel platforms are both Linux clusters. There is a project underway to port the codes that have traditionally run on the Tru64 platforms to the Linux clusters. The two-dimensional legacy codes will run on these nodes as well as the ASCI application mesh generation codes. These platforms will still generate much smaller data sets, with many more users, when compared to ASCI platform standards. Use of the file bundling tool, HTAR, will help insure that adequate throughput is delivered concurrently to multiple users with data sets made up of small files.

The LLNL, Tri-Lab and capacity computing usage scenarios outlined above are anticipated to be the predominant modes of operation for the code runs in FY02 but there will be others. For instance, some users will want to move their full data set immediately to the visualization server or HPSS after computations complete on the ASCI computer. From this it can be determined that the paths from the ASCI machine to the archive, from the ASCI machine to the Visualization server and the Visualization server to the archive will carry the preponderance of the network traffic. This is highlighted on the transfer rate diagrams given in later sections by showing these paths in red.

SCF Throughput Requirements

There are 8 major end points on the SCF network (the White machine, the SKY machine, EDTV, the Compaq cluster, the two new Linux clusters, the Visualization machines, and the Data Archive platform). Bandwidth estimates for all eight endpoints would need to be calculated to determine true network bandwidth requirements. The power and capacity of the ASCI platforms generate bandwidth requirements that is at least 4x the requirements of the other sources on the network. Additionally, there is a complication this year in that we don't actually know how powerful the EDTV system will be. For the purposes of this document we are estimating that EDTV will be comparable to White in terms of its salient characteristics. Because of the uncertainty surrounding the final capability of EDTV and because it will be operational for less than half the year, an upper bound of twice the capability of White was used to determine SCF throughput requirements. The methods used to estimate bandwidth requirements from the ASCI machines to the archive are listed in Table 2.

There is the issue of whether sustained throughput or peak throughput is being discussed. Peak throughput is defined as the maximum file transfer rate (in bytes per second) over the network from any one machine (source) to another machine (sink). Peak aggregate throughput is the maximum number of bytes per second that can be pushed from all sources to all sinks in the network. Sustained aggregate throughput is estimated to be 10% of peak aggregate throughput. This estimation of sustained aggregate throughput has been verified through LLNL user experiences. The transfer rate estimates in the table are peak aggregate throughput estimates.

Method	Bandwidth
Maximum bandwidth off White and EDTV Clusters	1,960 MB/s
User simulation scenario - 60TB of data in 24 hours	694 MB/s
10% of achieved disk cache transfer rate	620 MB/s
200 memory dumps per year x 10	800 MB/s
Disk cache saved 4 times per year x 10	430 MB/s
B Division Code Scenario: Move 28.8 TB in 15.5 hours	516 MB/s

Table 2: White/EDTV to Archive Aggregate Peak Transfer Rate Estimates

The "Maximum bandwidth off White and EDTV Clusters" estimate is based on sustaining the maximum amount of data transfer off these machines simultaneously. White has four Login nodes with 240 MB/s delivered on each. EDTV has four Login nodes with 250 MB/s delivered on each. The "User simulation scenario" estimate is based on moving 30TB of data in 24 hours from both machines to the archive simultaneously, for a total of 60TB of data transferred. This is based directly on a user simulation scenario.

The 10% of achieved disk cache transfer rate estimate is based on the model that applications will spend 10% of the time writing I/O to the local disk cache and 90% of the time computing the next I/O dump. Given this model, in order to pipeline data off the disk cache to storage before the next disk cache I/O dump the archive transfer rate needs to be approximately 10% of the disk cache rate. The estimate for the White and EDTV machines achieved disk cache transfer rate is based on achieving 25% of the rated disk cache bandwidth.

The third estimation method is based on storing 200 times the platform's memory size per year and is the method used in the FY00 I/O requirements document and in Appendix A of the FY01 I/O Blueprint. It is included here for year-to-year consistency. It should be noted that LANL does projections in this manner. However, LANL uses a factor of 750. LLNL chooses the 200x-scaling factor (times system utilization) to match the storage pressure seen in production during the 1HCY98. From these "sustained" transfer rate estimates a "peak" estimate was obtained (peak = 10 x sustained).

The fourth estimation method (disk cache saved 4 times per year X 10) is based on usage history of the SKY machine. The SKY disk cache became close to full after six months of operation. This was a new machine, which was only partially utilized, especially in the first few months of operation. Once the codes are optimized for the machine and the machine is functioning at high utilization, it is reasonable to assume that users will generate data at a rate that will fill the disk cache every 3 months. The memory dump and disk cache estimates use the amount of data and time frames specified to calculate a sustained rate. The multiplier of ten changes the estimate from a sustained rate to a peak rate.

The fifth estimation method is based on a B Division simulation code scenario.

The transfer rate requirements from the visualization servers (Tidalwave, Edgewater, and Whitecap) to the archive have not changed from last FY. The read rate remains at 20MB/s and the write rate at 70MB/s. The read rate is limited by our archive and is acceptable since reads of large data are rarely done. The write rate is based on the requirement to move a 6TB data set from the visualization servers to the archive in one day.

Method	Bandwidth
Move 1 dataset (6TB) from vis to archive in 24 hours	70 MB/s

Table 3: Visualization Server Bandwidth Estimates

The estimates in Table 4 document the additional bandwidth needed for data analysis work in SCF instead of data transfer to the archive. There are two usage modes we see today. The first is when the user moves data from a few timesteps at a time to the visualization servers for rendering. The second is when the VIEWS nodes on White are used for data reduction and then the reduced data set is moved to the visualization servers for rendering. The applications that are capable of this mode are EnSight and LLNL's VisIt. The bandwidth estimates from these two usage models are in the first two rows of the table below. The first estimation is based on moving 2 to 3 timesteps, 200 to 300 GBs each, to the visualization server in one hour.

The second estimation method is new this year and reflects this second usage mode we now see. Eric Brugger from B Program and the lead of the VisIt effort estimates that one interactive session using VisIt on White will require bursty 125MB/s of bandwidth from White to the visualization servers. We assume that two such interactive sessions will be the maximum.

This year we are planning on deploying one IBM 9 Mpixel flat panel. This panel sits behind a IBM Scalable Graphics Engine that has 8 GE inputs and 4 DVI outputs. The 4 DVI outputs drive the flat panel. We plan to do testing to determine the required number of GE inputs to drive this panel, but until then we can only say that it may take anywhere from 2 GEs to 8 GEs. Using a maximum of 60MB/s per GE link we require anywhere from 120MS/s to 480MB/s. We assume like the VisIt usage mode that this bandwidth is bursty. If this flat panel is successful we may be asked by our customers to deploy more in the future.

Method	Bandwidth
Move 1TB from White or EDTV to Vis in 1 hour	280 MB/s
Support two interactive VisIt data analysis sessions	250 MB/s
Image delivery from White to one IBM Scalable Graphics Engine located in either B-132 or B-111	120MB/s to 480MB/s

Table 4: Other Bandwidth Estimates to Support Data Analysis Services

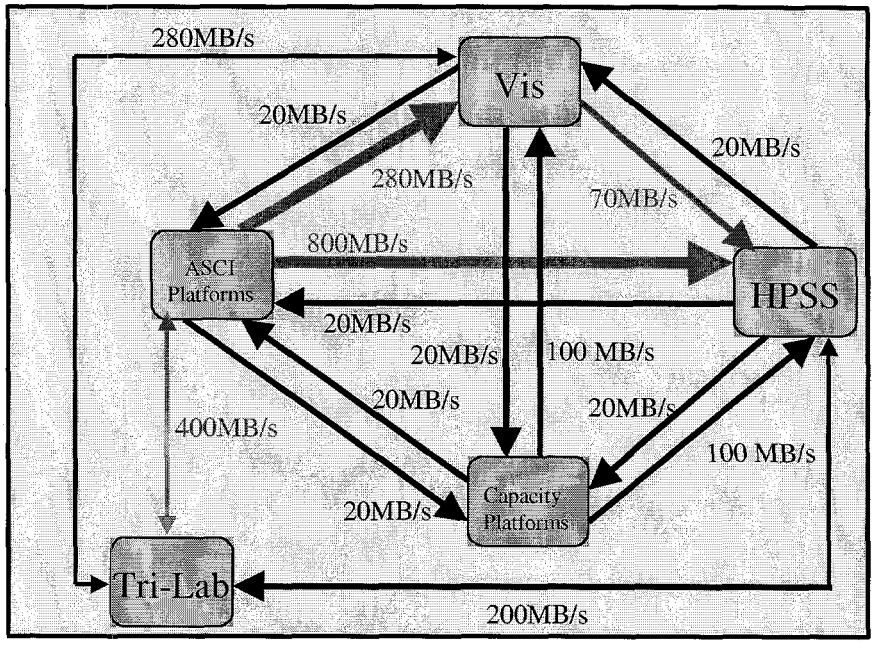

Figure 1. SCF Throughput Requirements

Figure 1 represents the throughput requirements associated with each major pipe in the SCF I/O network. It is important to note that the transfer rate requirements listed on figure 1 are specified to be "concurrent", i.e. the network infrastructure and end nodes must be capable of delivering all the specified link rates concurrently. There was almost universal consensus among the authors that we don't have users and applications that drive the full set of ASCI nodes to their maximum off-node I/O capabilities as is required by the 1960 MB/s estimate for the archive link, so the next highest estimate of 800 MB/s was chosen. This aggressive throughput requirement was chosen because bandwidth is of paramount importance to provide successful sinking of ASCI data. Visualization server transfer rate requirements were taken directly from the Terascale Assimilation Version 0.4 document as described in Table 3.

The links to and from the Tri-Labs deserve discussion. These links are for Sandia and LANL users running their codes on the LLNL ASCI machines and then storing to filesystems at their home site. The bandwidth required is 400 MB/s. This is half of what is required locally, but represents either half the local bandwidth of both White and EDTV, or the full local bandwidth requirement off of either White or EDTV, but not both. A full 1200 MB/s of throughput is required on the Tri-Lab link and the local HPSS link simultaneously.

The throughput requirements to HPSS from all sources total 970 MB/s. This requires that the HPSS disk cache be capable of sinking and sourcing data at a combined rate of 1940 MB/s. The disk cache must sink 970 MB/s of data while at the same time it must be able to migrate 970 MB/s of data to tape.

The link from the Capacity machines to the archive is specified at 100 MB/s. As the main capacity engine this is where the legacy codes and ASCI application mesh generation will be occurring. The remaining links between sources and sinks are all 20 MB/s. This number was chosen to indicate that heavy usage is not expected, but connectivity and reasonable throughput must be available.

SCF Aggregate Transfer Rate Requirements (MB/s)			
System	Write	Read	Total
ASCI	1500	80	1580
Vis	110	400	510
HPSS	260	1170	1430
Capacity	120	60	180

Table 5. FY02 SCF end-node aggregate transfer rate requirements

SCF Archive Capacity Requirements

As has already been stated in this document, the ASCI platform archive requirements dwarf those of other platforms. This being true, only ASCI machines are being used to calculate archive capacity requirements. One can only make educated guesses about the final computing power or configuration of EDTV. What is absolutely certain is that SKY and White will provide 16.2 TFLOPs of capability for all of FY02. Leaving SKY out of the calculations and assuming that EDTV will have exactly the same characteristics as White makes calculations simple. It is a conjecture that there will be 24.6 TFLOPs of computing capability on the floor the entire year for a plus up of 8.4 TFLOPs. This is the guess that was made because it does not sacrifice reason for the sake of simplicity, but again, it is simple. Two methods were used to estimate FY02 archive capacity requirements. The first method is based on FY01 storage history. The peak aggregate bandwidth experienced from ASCI platforms to the archive was approximately 200 MB/s. The total number of bytes stored in FY01 was approximately 200 TB. This implies that the average transfer rate to the archive was 3.5% of the peak aggregate throughput. The capacity required if. Given 800 MB/s aggregate bandwidth to storage at peak conditions in FY02, the equation for required capacity if 3.5% of peak bandwidth to the archive is stored is 28 MB/s x 31.5 million per year. This method yields an estimate of 882 Terabytes of data stored in one year's time.

The second method used to estimate capacity requirements is based on input from users working toward programmatic milestones. Users on the Capacity Clusters are working on 2-dimensional problems aimed at meeting current verification objectives and some small 3-D problems. These applications generate less than a few percent of the data created by the ASCI 3-dimensional applications. The requirements of the ASCI codes, as perceived by a few code developers are being used to determine this estimate.

The A and B Division code developers asked us to estimate next years use based on applying a few factors to the use of the data archive in FY01. They suggested we estimate this year's requirements by doubling the amount of data stored last year because of the increased platform capability and then doubling that number again in case throughput has improved enough to encourage increased use of the archive. We used a similar estimation method last year and estimated that we would go from 50TB stored during FY00 to 300TB stored during FY01. In fact, we stored approximately 200TB in FY01. So the estimation technique was not terrible and more importantly it was high, rather than low. If we double the 200TB stored in FY01 and then double it again to account for increased use because of increased throughput, we estimate that 800TB of data will be transferred into the SCF archive in the next year. We have approximately 300TB of data currently stored in the SCF data archive, which implies that we must

have at least 1.1PB of capacity to store next year's data. This number is an upper bound on what our users have estimated.

Table 6 summarizes the capacity estimates. The user estimate for data that will be archived is less than our projection. What has definitely been true since the ASCI machines have been on the floor is that users have not stored most of the data they have generated. This is not surprising given that the runs that are being done on the ASCI machines are code confirmation runs rather than true certification runs. Once designers have accepted and adopted the 3-D codes, we can expect that a much greater percent of the data generated during a run will be stored.

Method	Capacity
4/6% of peak bandwidth to archive	1.2 PB
User estimate of data to be archive (A and B Division)	1.1 PB

Table 6: Summary of SCF capacity estimates

From these estimates a capacity requirement of 1.1 PB was selected. We currently have 1.1PB of tape media available to populate the SCF storage system. We have 300 TB of data stored in the silo, so we have 800 TB of media available. The silos are currently capable of holding 2.7 PB of data if all slots are populated with 90 GB/cartridge media.

FY01 SCF Storage Capacity Requirement (PB)			
System	*Start*	*End*	*Delta*
SCF	**1.1**	**1.1**	**0**

Table 7: FY02 SCF Capacity Requirement

FY02 OCF Computing Platform Changes

There are several changes anticipated for the OCF in FY02. First, the Linux development project will be building large temporary clusters for development of cluster tools and global file systems (e.g., PvFS or GFS). It is not anticipated that these resources will place much demand on the archive or visualization resources other than testing. In addition, as mentioned in the SCF Computing Platform Changes section, we anticipate having the PCR P4A (890 GFLOP/s) cluster in the OCF for a period of at least three months (1QFY02) doing science runs. Given that this cluster has a great deal of local disk space (10.4 TB) and very little global file space (7.0 TB), and high sustained parallel FTP rates off the cluster (120 MB/s from each of two Login nodes for a total of 240 MB/s), it is anticipated that the migration of files to the OCF visualization and archive resources will be substantial. In addition, there is significant institutional pressure for a much larger Linux cluster in FY02 to support a large set of science runs. Several budget scenarios are being discussed, but the FY02 OCF Linux cluster will likely be in the 4-8 TFLOP/s range and have 2-4 TB of memory and 100-200 TB of global disk and be able to support 4 Login nodes with 250 MB/s sustained parallel FTP performance each for a total of 1.0 GB/s. This Linux OCF Capability Resource (LOCR) will probably be connected via NTON to San Diego Supercomputer Center (SDSC) and NERSC and utilized as part of a distributed computing resource. San Diego's emphasis will be on "Big Data." There will be a huge demand for movies and data analysis coming out of these OCF science runs.

Table 8 below lists the anticipated FY02 OCF machine characteristics of interest for the purposes of bandwidth and capacity requirements determination.

Machine	Compute Power	Memory Capacity	Disk Cache Bandwidth	Disk Cache Capacity	External I/O
Frost (ASCI)	1.6 TFLOP/s	1.0 TB	1.6 GB/s	40 TB	240 MB/s
Blue (ASCI)	701 GFLOP/s	384 GB	2.2 GB/s	20 TB	180 MB/s
Tera Cluster 2000 (M&IC)	683 GFLOP/s	280 GB	270 MB/s	9 TB	30 MB/s
Vis Platforms	76 GFLOP/s	148 GB	800 MB/s	25 TB	20 GbEnet ???
LOCR Clusters (ASCI)	4-8 TFLOP/s	2 TB	6 GB/s	100 TB Global	1.0 GB/s

Table 8: FY02 OCF platform configurations

OCF Throughput Requirements

As described above, the Open Computing Facility (OCF) computing platforms have evolved significantly over the past few years. Blue Light will change things dramatically when it is added to the mix, but is not planned for until 2004.

Table 9 estimates archive bandwidth requirements for the platforms listed in Table 7 coalesced into two functional groups: ASCI platforms and Multiprogrammatic and Institutional Computing platforms. Note that grouping these machines by funding stream is not exact, that is, both ASCI and M&IC funds may have been used to buy these platforms. The estimation methods used for OCF throughput requirements are two of those used to estimate SCF throughput requirements. The 10% of achieved disk cache transfer rate estimate is based on the model that applications will spend 10% of the time writing I/O to the local disk cache and 90% of the time computing the next I/O dump. Given this model, in order to pipeline data off the disk cache to storage before the next disk cache I/O dump the archive transfer rate needs to be approximately 10% of the achieved disk cache rate. The estimate for the ASCI and M&IC platforms achieved disk cache transfer rate is based on achieving 25% of the rated disk cache bandwidth.

The next estimation method is based on storing 200 times the platform's memory size per year and is the method used in the FY00 I/O requirements document and in Appendix A of the FY01 I/O Blueprint. It is included here for year-to-year consistency. It should be noted that LANL does projections in this manner. However, LANL uses a factor of 750. LLNL chooses the 200x-scaling factor (times system utilization) to match the storage pressure seen in production during the 1HCY98. From these "sustained" transfer rate estimates a "peak" estimate was obtained (peak = 10 x sustained).

Table 9 estimates archive bandwidth requirements for the platforms in Table 8 coalesced into two functional groups: ASCI platforms and Multiprogrammtic and Institutional Computing platforms. The estimation methods used for OCF throughput requirements are two of those used to estimate SCF throughput requirements.

Method	M&IC Transfer Rate	ASCI Transfer Rate
10% of achieved disk cache transfer rate	6.7 MB/s	245 MB/s
200 memory dumps per year x 10	18 MB/s	219 MB/s

Table 9: OCF Data Sources to Archive Bandwidth Estimates

Figure 2 represents the throughput requirements associated with each link in the OCF I/O network. The ASCI systems transfer rate to archive requirement is 200MB/s. Although it is less than both estimates above, it represents more than double the throughput increase available today and reflects the fact that the LOCR clusters will not be online the entire year. The M&IC cluster transfer rate requirement is 50 MB/s. Traditionally the work on the M&IC clusters consists of a large number of concurrent users generating moderate amounts of data. Since we now have a file bundling tool, throughput is needed to accommodate large data sets as well as sufficient aggregate throughput to insure each user's transfer rate is reasonable.

The transfer rate from the Visualization server to the archive is specified at 50 MB/s. The estimation techniques applied to the ASCI and M&IC machines are not applied to the Visualization server since these techniques are geared for general computing platforms. In general, data produced on the ASCI platform is transferred to the Visualization server, processed (typically resulting in reduced data) and if the results are interesting/important they are stored in the archive. The 50 MB/s requirement is aimed at providing enough bandwidth to accommodate this pipeline usage.

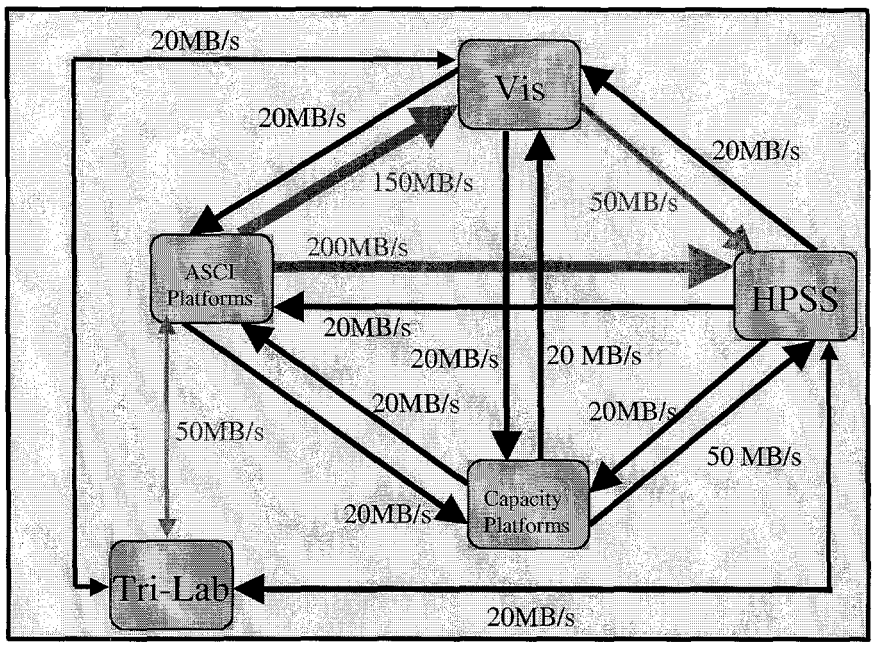

Figure 2. OCF Throughput Requirements

OCF Aggregate Transfer Rate Requirements (MB/s)			
System	Write	Read	Total
ASCI	420	110	530
Vis	110	190	300
HPSS	80	320	400
Capacity	90	60	150

Table 10. FY02 OCF end-node aggregate transfer rate requirements

OCF Capacity Requirements

There were no discussions with users on OCF storage capacity requirements. There is considerable uncertainty concerning the requirements of remote users, in particular ASCI university alliances. Since the capacity estimate is based on transfer rates, which in turn are based on the specifications of the computational engines (memory size and disk cache bandwidth) this estimate provides an upper bound based on the capabilities of the compute engines.

The estimate for OCF archive capacity is based on the capacity required if 3.5% of peak bandwidth to the archive is stored. Given 300 MB/s aggregate bandwidth to storage at peak conditions, the equation is 10.5 MB/s x 31.5 million seconds per year. This method yields an estimate of 330 terabytes of additional required capacity.

The OCF storage system currently has 111TB stored of a total 650 TB populated tape capacity. The additional 330 TB of estimated FY02 capacity is well within the current 650TB of purchased capacity, so there is no requirement to purchase additional media for the OCF archive. The total capacity of the OCF tape archives, if the remaining silo slots were fully populated with high capacity (90 GB/cartridge) tape cartridges is 2.16 PB.

FY00 OCF Storage Capacity Requirement (TB)			
System	*Start*	*End*	*Delta*
OCF	650	650	0

Table 11: FY02 OCF Capacity Requirement

Appendix B: Procurements – *Procurement Sensitive Information*

Removed for Public Release Version

Appendix C: Schedule of Blueprint Deliverables

The following is a schedule of all deliverables outlined in this Blueprint. The deliverables outlined in the Blueprint are directly supported by work in the ASCI VIEWS, PSE, PC and DisCom2 Program FY02 Implementation Plans. Note that the quarters are for calendar years rather than fiscal years.

ID	Task Name	Duration
1	**Issue A: High Archive Costs to Support Future Platforms**	**165 days**
2	**Improved IBM Mover Platform**	**75 days**
3	Develop IBM M80 test plan	10 days
4	Test M80	45 days
5	IBM M80 testing report	20 days
6	**HPSS SAN Strategy Investigation**	155 days
7	**Investigate mover running on ASCI platform node**	**125 days**
8	Write test plan	20 days
9	Install/test HPSS Local File Mover	5 days
10	Install/test FTP client to HPSS Local File Mover	5 days
11	Borrow SAN switch	15 days
12	Install/test SAN switch connected to HBA & tape devices	15 days
13	Execute test plan	30 days
14	Cost/benefit analysis and recommendation	20 days
15	**Investigate low cost LINUX mover platform**	**165 days**
16	Write test plan	5 days
17	Port HPSS mover to Intel Linux box	120 days
18	Execute test plan	15 days
19	Cost/benefit analysis and recommendation	20 days
20	**Investigate SAN fabric technology**	60 days
21	Recommend and Procure SAN fabric technology	15 days
22	**Issue B: Small File Sizes**	**162 days**
23	Target users for adoption of HTAR	120 days
24	Deliver report of user target results	10 days
25	Machine-machine HTAR	120 days
26	HTAR ported to LINUX	30 days
27	**ISSUE C: Lack of RAIT Capability**	**26 days**
28	Meet with tri-Lab	5 days
29	Meet with vendors	5 days
30	Recommend a go-forward strategy	15 days
31	**Issue D: Disaster Recovery**	**83 days**
32	Meet with tri-Lab	2 days
33	Meet with users	20 days
34	Write report	20 days
35	**Issue E: High Speed Archive Throughput**	**230 days**
36	Procure more HPSS disk cache	30 days
37	Phase out IBM3494	230 days
38	Expand tape stripe width behind HTAR	10 days
39	**Issue F: Grid Services**	**300 days**
40	Evaluate grid services client for work on Q	300 days
41	**Issue G: Data Movement**	**365 days**
42	Mover protocol changes in HPSS	365 days
43	**Issue H: Remote Visualization**	**120 days**
44	Bring whitecap online	120 days
45	**Issue I: Usage and Usage Satistics**	**120 days**
46	Revamp usage statistics gathering	120 days
47	**Issue K: Platform to Vis File Transfers**	**40 days**
48	Test NFT from white to SGIs and publish performance	5 days
49	Periodically monitor and publish vis data xfr performance	40 days
50	**Issue L: Network Performance Stability and Configuration**	**181 days**
51	Channel bonding investigation	20 days
52	Channel bonding report delivered	20 days
53	Gigabit ethernet flow control investigation	20 days
54	Gigabit ethernet flow control report delivered	20 days
55	10GigE deployed between B451 and B113	30 days
56	Testing of OC48 Ultrfastlanes	120 days